CLAIMS

What is claimed is:

1. A method of determining if a query document matches one or more documents in a database, the method comprising:

generating a bit profile of the query document based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

comparing the bit profile of the query document against bit profiles associated with a first plurality of documents from the database to determine if the query document matches one or more of the first plurality of documents.

2. The method of claim 1 further comprising:

performing spectral analysis on the bit profile of the query document to determine global statistics of the query document; and

comparing the global statistics of the query document against global statistics associated with a second plurality of documents from the database to identify the first plurality of documents.

3. The method of claim 2 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of a dominant line spacing in the query document.

4. The method of claim 2 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation

of a proportion of the query document that is text.

5. The method of claim 2 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of a location of text in the query document.

6. The method of claim 2 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of text concentration in the document, the estimation of text concentration indicating a lengthwise measure of a proportion of the query document that is text.

7. The method of claim 1 further comprising precomputing the bit files associated with the first plurality of documents and storing the precomputed bit profiles in the database.

8. The method of claim 1 wherein comparing the bit profile of the query document against bit profiles associated with the first plurality of documents comprises cross correlating the bit profile of the query document against the bit profiles associated with the first plurality of documents from the database.

9. The method of claim 8 wherein cross correlating the bit profile of the query document against the bit profiles associated with the first plurality of documents from the database comprises generating respective vector products of the bit profile of the query document the bit profiles associated with the first plurality of documents from the database.

10. The method of claim 9 wherein the query document is determined to match one or more of the first plurality of documents for which the respective vector product exceeds a threshold.

11. A method of determining if a query document matches one or more documents in a database, the method comprising:

identifying up endpoints and down endpoints in the query document, the up endpoints representing tops of features in the query document and the down endpoints representing bottoms of features in the query document;

generating a set of descriptors for the query document based on locations of the up endpoints and the down endpoints; and

comparing the set of descriptors for the query document against respective sets of descriptors associated with the one or more documents in the database to determine if the query document matches at least one of the one or more documents.

12. The method of claim 11 wherein generating a set of descriptors for the query document based on locations of the up endpoints and the down endpoints comprises

identifying text lines in the query document based on concentrations of up endpoints and down endpoints along scanlines of the query document; and

generating the set of descriptors based on distances between selected up endpoints and selected down endpoints within the text lines in the

query document.

13. The method of claim 12 wherein identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document comprises:

determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines; and

identifying respective pairs of scanlines that have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines.

14. The method of claim 11 wherein the query document is in a compressed form in which respective runs of pixels are encoded in one of a plurality of encoding modes, and wherein identification of the up endpoints and down endpoints is unaffected by the encoding mode.

15. A method of determining if a query document matches one or more documents in a database, the method comprising:

generating a bit profile of the query document based on the number of bits required to encode each of a plurality of rows of pixels in the query document;

comparing the bit profile of the query document against bit profiles associated with a first plurality of documents from the database to identify one or more candidate documents;

identifying endpoint features in the query document;

generating a set of descriptors for the query document based on locations

of the endpoint features; and

comparing the set of descriptors for the query document against respective sets of descriptors for the one or more candidate documents to determine if the query document matches at least one of the one or more candidate documents.

16. The method of claim 15 further comprising

performing spectral analysis on the bit profile of the query document to determine global statistics of the query document; and

comparing the global statistics of the query document against global statistics associated with a second plurality of documents from the database to identify the first plurality of documents, the first plurality of documents being a subset of the second plurality of documents.

17. The method of claim 16 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of at least one of a dominant line spacing in the query document, a proportion of the query document that is text, a location of text in the query document, and a text concentration.

18. The method of claim 15 wherein comparing the bit profile of the query document against bit profiles associated with the first plurality of documents comprises cross correlating the bit profile of the query document against the bit profiles associated with the first plurality of documents from the database.

19. A method of generating a set of descriptors for identifying a document, the method comprising:

identifying up endpoints and down endpoints in the document, the up endpoints representing tops of features in the document and the down endpoints representing bottoms of features in the document;

identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document; and

generating a set of descriptors based on distances between selected up endpoints and selected down endpoints in the concentrations of up endpoints and down endpoints.

20. The method of claim 19 wherein identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document comprises:

determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines; and

identifying respective pairs of scanlines that have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines.

21. The method of claim 19 wherein identifying text lines in the document based on concentrations of up endpoints and down endpoints along scanlines of the document comprises:

determining a dominant line spacing in the document;

determining the number of up endpoints and the number of down endpoints that lie on each of the scanlines; and

identifying as text lines respective scanline pairs in which the constituent scanlines are separated by a distance less than the dominant line spacing and in which the constituent scanlines respectively have a local maximum number of up endpoints and a local maximum number of down endpoints as text lines.

22. The method of claim 21 wherein the dominant line spacing is determined based on spectral analysis of locations of the endpoints in the document.

23. The method of claim 19 further comprising generating a respective endpoint profile for each of the scanlines, the endpoint profile including a count of up endpoints identified on the scanline and a count of down endpoints identified on the scanline, and wherein identifying text lines based on concentrations of up endpoints and down endpoints along scanlines of the document comprises reducing all but local maximums of the counts of up endpoints and the counts of down endpoints in respective endpoint profiles.

24. The method of claim 19 wherein identifying text lines based on concentrations of up endpoints and down endpoints along scanlines of the document comprises:

generating a count of up endpoints and a count of down endpoints for each of the scanlines;

identifying a first scanline within a locality of scanlines that has the

highest count of up endpoints;

reducing the count of up endpoints associated with each scanline within the locality of scanlines except the first scanline;

identifying a second scanline within the locality of scanlines that has the highest count of down endpoints; and

reducing the count of down endpoints associated with each scanline within the locality of scanlines except the second scanline.

25. The method of claim 24 wherein identifying the first scanline within the locality of scanlines that has the highest count of up endpoints comprises:

determining a dominant line spacing of the document; and

defining the locality of scanlines to be scanlines within a range greater than the dominant line spacing but less than twice the dominant line spacing.

26. The method of claim 19 wherein generating a set of descriptors based on distances between selected up endpoints and selected down endpoints comprises defining an ascender zone and a descender zone for each of the text lines, the selected up endpoints being up endpoints in the ascender zone and the selected down endpoints being down endpoints in the descender zone.

27. The method of claim 26 wherein defining an ascender zone and a descender zone for each of the text lines comprises:

defining a region above an x-height line of a first text line of the text

lines to be the ascender zone for the first text line; and

defining a region below the baseline of the first text line to be the descender zone for the first text line.

28. The method of claim 27 wherein the ascender zone of the first text line is bounded in part by the descender zone for the preceding text line.

29 The method of claim 19 wherein generating a set of descriptors based on distances between selected up endpoints and selected down endpoints comprises generating for a first text line of the text lines a first descriptor that includes a plurality of distance measurements, each distance measurement indicating a distance between a reference point and a respective endpoint of the selected up endpoints and the selected down endpoints.

30. The method of claim 29 wherein the reference point is one of the selected up endpoints and the selected down endpoints.

31. The method of claim 29 wherein each distance measurement indicating the distance between the reference point and the respective endpoint is a relative distance to another endpoint of the selected up endpoints and the selected down endpoints.

32. The method of claim 19 wherein the document is in a compressed form in which respective runs of pixels are encoded in one of a plurality of encoding modes, and wherein identification of the up endpoints and down endpoints is unaffected by the encoding mode.

33. The method of claim 19 wherein the document has been compressed using Group 4 compression.

34. A method of generating information that can be used to identify a document, the method comprising:

generating a bit profile based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

performing spectral analysis on the bit profile to determine global statistics of the document.

35. The method of claim 34 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of a dominant line spacing in the document.

36. The method of claim 35 wherein generating an estimation of a dominant line spacing comprises generating a power spectrum density from the bit profile and calculating the estimation of the dominant line spacing from a peak value in the power spectrum density.

37. The method of claim 34 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of a proportion of the document that is text.

38. The method of claim 37 wherein generating an estimation of a proportion of the document that is text comprises generating a power spectrum density from the bit profile and calculating the estimation of

the proportion of the document based on an energy under a peak value in the power spectrum density.

39. The method of claim 34 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of a location of text in the document.

40. The method of claim 39 wherein generating an estimation of a location of text in the document comprises:

applying a bandpass filter to the bit profile to generate a text energy profile; and

determining a centroid of the text energy profile to be the estimation of the location of text in the document.

41. The method of claim 40 wherein applying a bandpass filter to the bit profile comprises:

determining a dominant line spacing frequency of the document; and

selecting a center frequency of the bandpass filter based on the dominant line spacing frequency.

42. The method of claim 34 wherein performing spectral analysis on the bit profile to determine global statistics comprises generating an estimation of text concentration in the document, the estimation of text concentration indicating a lengthwise measure of a proportion of the document that is text.

43. The method of claim 42 wherein generating an estimation of text concentration in the document comprises:

applying a bandpass filter to the bit profile to generate a text energy profile; and

determining the estimation of the text concentration based on a length of the text energy profile.

44. An article of manufacture including one or more computer-readable media that embody a program of instructions to configure a processing system to determine if a query document matches one or more documents in a database, wherein the program of instructions, when executed by one or more processors in the processing system, causes the one or more processors to:

generate a bit profile of the query document based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

compare the bit profile of the query document against bit profiles associated with a first plurality of documents from the database to determine if the query document matches one or more of the first plurality of documents.

45. The article of claim 44 wherein the one or more computer-readable media include one or more non-volatile storage devices.

46. The article of claim 44 wherein the one or more computer-readable

media include a propagated data signal.

47. An article of manufacture including one or more computer-readable media that embody a program of instructions to configure a processing system to determine if a query document matches one or more documents in a database, wherein the program of instructions, when executed by one or more processors in the processing system, causes the one or more processors to:

identify up endpoints and down endpoints in the query document, the up endpoints representing tops of features in the query document and the down endpoints representing bottoms of features in the query document;

generate a set of descriptors for the query document based on locations of the up endpoints and the down endpoints; and

compare the set of descriptors for the query document against respective sets of descriptors associated with the one or more documents in the database to determine if the query document matches at least one of the one or more documents.

48. An article of manufacture including one or more computer-readable media that embody a program of instructions to configure a processing system to determine if a query document matches one or more documents in a database, wherein the program of instructions, when executed by one or more processors in the processing system, causes the one or more processors to:

generate a bit profile of the query document based on the number of bits

required to encode each of a plurality of rows of pixels in the query document;

compare the bit profile of the query document against bit profiles associated with a first plurality of documents from the database to identify one or more candidate documents;

identify endpoint features in the query document;

generate a set of descriptors for the query document based on locations of the endpoint features; and

compare the set of descriptors for the query document against respective sets of descriptors for the one or more candidate documents to determine if the query document matches at least one of the one or more candidate documents.

49. A data processing system comprising:

a database of document images; and

a computer that includes a processing unit and a memory, the memory having stored therein a program of instructions to configure the computer to determine if a query document matches one or more documents in the database, wherein the program of instructions, when executed by the processing unit of the computer, causes the computer to:

generate a bit profile of the query document based on the number of bits required to encode each of a plurality of rows of pixels in the document; and

compare the bit profile of the query document against bit profiles associated with a first plurality of documents from the

database to determine if the query document matches one or more of the first plurality of documents.

50. A data processing system comprising:

a database of document images; and

a computer that includes a processing unit and a memory, the memory having stored therein a program of instructions to configure the computer to determine if a query document matches one or more documents in the database, wherein the program of instructions, when executed by the processing unit of the computer, causes the computer to:

identify up endpoints and down endpoints in the query document, the up endpoints representing tops of features in the query document and the down endpoints representing bottoms of features in the query document;

generate a set of descriptors for the query document based on locations of the up endpoints and the down endpoints; and

compare the set of descriptors for the query document against respective sets of descriptors associated with the one or more documents in the database to determine if the query document matches at least one of the one or more documents.

51. A data processing system comprising:

a database of document images; and

a computer that includes a processing unit and a memory, the memory having stored therein a program of instructions to configure the

computer to determine if a query document matches one or more documents in the database, wherein the program of instructions, when executed by the processing unit of the computer, causes the computer to:

generate a bit profile of the query document based on the number of bits required to encode each of a plurality of rows of pixels in the query document;

compare the bit profile of the query document against bit profiles associated with a first plurality of documents from the database to identify one or more candidate documents;

identify endpoint features in the query document;

generate a set of descriptors for the query document based on locations of the endpoint features; and

compare the set of descriptors for the query document against respective sets of descriptors for the one or more candidate documents to determine if the query document matches at least one of the one or more candidate documents.